新聞記事における書き出し文の構造
－複文を中心に－

乾　裕子[1]**　木田敦子[1]　橋本三奈子[2]　落谷　亮[3]　西野文人[3]
[1] {hinui,akida}@ibs.or.jp　[2] hashimoto@aisys.se.fujitsu.co.jp
[3] {ochi,nisino}@flab.fujitsu.co.jp
[1]計量計画研究所　[2]富士通　[3]富士通研究所

1．はじめに

情報抽出では、文章のスタイルが定型的である新聞記事を対象とすることが多い。特に書き出し文は、一般にいつ・誰が・どこで、何を、どうしたといった5W1H型の固定的な表現で記述される。現在、われわれが開発を進めているシステムでも、この性質に着目して情報抽出を行っている。しかし、書き出し文の中には、名詞句を修飾する連体修飾節や述語を修飾する連用節を含む複文のように定型的でない表現が表れる。

1.1 事象モデルに基づく情報抽出システム

本研究で開発されている情報抽出システムは、新聞記事テキストから製品販売、組織合併、研究開発、業務提携、新事業への参入といった企業動向を事象構造（誰が、いつ、どこで、何を、など）として抽出する[5][6][7]。表現を解析して内容を理解するのではなく、シナリオに基づいて個々の属性値とテキスト中の表現との対応関係をとっている。シナリオとは企業が新製品を発売する、ある親企業が子会社同士を合併させるといった事象を指し、属性値はシナリオに想定される企業名（発売元、親会社、子会社など）や企業所在地・社長名などの関連情報、製品名、日付などを指す。このシナリオテンプレートに基づくトップダウンな手法は、形態素解析によるボトムアップな手法に比べて、辞書に登録されていない製品名や組織名を抽出できるという利点がある。

1.2 行為表現に着目した事象の抽出

シナリオ中の個々の属性値は、事象が特定されてから抽出される。したがって、文末表現を中心とした事象判定が重要になる[6]。この事象判定は、文末述語を中心とした行為表現に着目して行っている。そのため「実用化に向けて開発するとともに、製品化を進める」「A社がB社と合併し、新分野に乗り出す」のように複数の事象が表れるすなわち複数の述語が表れる複文の事象判定は困難である。

本研究では、1）複文として記述される事象にはどのようなものがあるか、2）事象間の関係はどうなっているかを明らかにし、出現する事象のパタンを調べ、複数の事象を取り扱うシナリオを作成するために注意すべき点を洗い出すことが目的である。したがって、問題解決の観点として、1では事象にかかわる述語表現に、2では接続表現に着目して分析を進める。

2．複文からの情報抽出

一般に、述語が複数含まれた文を複文と呼ぶ。動詞や形容詞の中止形（連用中止形、テ形中止形、体言止め）あるいは接続表現が末尾に表れた節、連体修飾節を含む文を指す[1][4]。本稿では、このうち接続表現の表れた複文を対象に分析する。また、複文を構成する複数事象が必ずしも主従関係ではないと考えるため、接続表現に前接するいわゆる従属節を前件、後続する部分を後件と呼ぶ。

複文を研究する際には前件と後件に表れる各事象間の　a）論理的関係、b）時間的、アスペクト的関係、c）同一指示関係を把握することが重要といわれる[3]。複文において同一指示関係とは、行為事象の主体や対象が前件と後件で同一のものを指すかどうかであり、情報抽出ではもっと

** 九州工業大学大学院情報科学専攻（h_inui@pluto.ai.kyutech.ac.jp）

も重要な問題である。したがって、分析の際には同一指示表現に着目して接続表現の関係を見直す。

3．複文における事象共起パタン

本節では、複数の事象に出現パタンがあるかどうかを調べるため、あるひとつの事象を取り出し、共起する事象について調べた結果を示す。

3.1 分析にむけての作業と結果

日刊工業新聞 401651 件の記事から、①連用中止形 ②テ形中止形 ③体言止め ④接続表現を含む文を取りだし複文データとする（91297 文）。接続表現は、複文に関する著述を参考に選定した[2][4]。尚、本稿に挙げる例文は、ことわりのない限り日刊工業新聞から取っている。

表１の開発事象データは、複文データから「開発スル（活用形含む）」を含む文を取りだしたものである。この中には、「持田製薬（社長持田英氏）が開発したβ型インターフェロンが米国の医薬研究機関で研究対象に取り上げられることになった。」といった連体修飾節に「開発」を含む約 5000 文も含まれている。これらは、分析の対象から除外する。

表１で、後件に表れる事象が前件の６倍以上あるのは現時点での前件・後件の決め方による。

例）日本酸素（社長野崎次郎氏）は、化合物半導体の製造装置を拡充するため、量産型ＭＯＣＶＤ（有機金属化学気相成長）装置およびＣＢＥ（ケミカル・ビーム・エピタキシー）装置を開発、販売を開始した。

上記の例で「～開発、」は、販売事象の前件としてのみ捉えており、「～拡充するための、」の後件述語として取り出していないためである。これを扱うのは今後の課題である。

表２は、開発事象における前件高頻度の「販売・発売」の特徴的な事象パタンを挙げたものである。パタンの具体例を下記に示す。

例）【宇都宮】日本エース（宇都宮市、社長赤羽輝一氏、電０２８６・３５・００８８）は、ツインブラシ機構を採用した新タイプの手動式掃除機「ＡＣＥＵＬＴＲＡ―２Ｗ＝写真」を開発、全国の代理店経由で販売を開始した。

開発スル、販売スルをまとめて開発・販売事象に「製品化」「商品化」、参入事象には「参入」のほかに「分野に進出する」「開拓に乗り出す」といった表現も含まれている。表３は、後件に開発事象が表れた際、前件事象として頻度の高い提携のパタンである。

表１　開発事象と共起する高頻度の事象

開発事象（12713文）	連体修飾節に表れる文（5507文） 中止形や接続表現を介して表れる文（7206文）		
後件事象（6278文）		前件事象（928文）	
事象	頻度	事象	頻度
販売・発売	3252	提携	48
受注	362	開発	23
出荷	319	課題・問題	20
実用・製品化	188	～化	16
納入	163	維持	11
試験・実験	116	向上	11
量産・生産	76	利用・使用	11
市場投入	66	注目さ	10
設立	60	問題・課題	9
成功	59	対応	7
導入	56	アップ	5
参入	52	拡大	4
稼働	42	強化	4
公開	40	研究開発	4
事業化	40	性能向上	4
提供	36	要請・需要	4
営業	34	支援	4
認可・承認	32	防止する	4
運用	31	改善	3
採用	27	確保	3
サービス	26	確率	3
開発	24	再利用	3
完成	23	ブーム	3
契約	21	指導	3

表２　販売・発売の事象パタン

前件	後件			
＊を開発、2881	＊発売	する。		1081
		した。		957
	＊販売	を	開始＊	216
			始め＊	45
		に	乗り出＊	54
＊を開発し、327	＊発売	する		151
		した		38
	＊販売	を	開始＊	56
			始め＊	17
		に	乗り出＊	17
＊を開発するため				7
＊を開発するとともに				7
：				
計				3257

表3　提携の事象パタン

前件	後件		
＊と提携＊	＊開発	する	。
＊と業務提携＊			ことになった
＊と技術提携＊			と発表する
＊と協力＊			と発表した
＊と手を組んで＊		した	。

例）ニチイ（社長小林敏峯氏）は小売業としては初めて蔵元と提携し、オリジナルの日本酒を開発した。

3.2 考察

3.1節では、事象の種類と共起のパタンを中心に見てきた。これらと事象間の関係とは実質的に切り離すことができないので、ここでは事象パタンごとの傾向と、表れた接続表現の組み合わせについて考察する。

表4　事象別の接続表現出現状況

事象（事象数）	開発 (12713)	合併 (382)	販売 (1717)
前件末に上記事象が表れる数→	6278	252	164
、	4485	6	72
し、	808	147	10
ため	363	17	56
とともに	105	11	10
が	42	5	6
ほか	6	2	4

前掲の表2からも推察できるように、開発事象を示す前件末尾の形式でもっとも多いのは体言止め（「開発、」）である（4485文）。このときの後件事象には「販売・発売」「受注」「出荷」「実用化・製品化」「量産・生産」など頻度の高いものが表れている。これらは、どの接続表現のときにも頻出するが、とくに「とともに」は共起のしかたが類似している。「ため」の場合は、後件に「設立」「提携」が多い。逆に後件に開発事象が表れるとき、「ため」の前件には「事業の拡大・強化」「精度向上」が表れる。「が」には一定の傾向が見られない。逆に考えると、事象間の関係の薄いことが特徴であるといえる。因果関係も時系列的関係も見られにくい事象が表れるため、単純な事象の併記と考えられる。

販売事象は比較的、開発事象と似た共起パタンを持つが、開発に比べて販売事象自体が後件に表れることが多い。これは、研究開発～製造～販売の流れを記述する際、時間的にもっとも最後に表れる事象だからであろう。

合併事象では新会社の開始や社名変更、新事業への参入が記述されている。このとき「ため」は、前件に表れることが少なく「事業の強化・拡大」の後件に表れる。

以上の分析から、新聞記事書き出し文の複文における事象の記述では接続表現と事象の共起パタンに相互関係が見られ、事象判定に向けて規則化できることがわかった。

4．接続表現による構造的・意味的特徴

事象間の関係を調べるにあたり、接続表現により関係が明示的である文を選ぶ。下記に挙げた例は、「事象の併記」という観点から見ると性質の共通した接続表現である。しかし、直観的にも「とともに」＞「ほか」＞「が」の順序で、事象の同時並行性が弱まると感じられる。

例）ティアック（社長谷勝馬氏）は二十五日、住友化学工業（社長森英雄氏）と共同で高解像度の光ディスク「ＭＡ―２５０Ｗ」を開発するとともに、同ディスクを活用し解像度と記録性を飛躍的に向上させた白黒タイプのレーザービデオディスクレコーダー「ＬＶ―２５０Ｈ＝写真」を製品化、販売活動を開始した。

山武ハネウエル（社長沖信春男氏）は温度調節計など計装機器分野で攻勢をかけるため、新たに通信機能を付けたデジタル指示調節計「ＳＤＣ２００」、デジタルプログラム調節計「ＤＣＰ２００＝写真」の二機種を開発したほか、従来の調節計、記録計にも通信機能を追加した。

秋田県工業技術センター（秋田市新屋町字砂奴寄四ノ一、所長松岡稔氏、電〇一八八（６２）三四一四）は射出成形金型におけるランナーバランスの最適条件を支援することを目的に、ランナーおよびキャビティー内の樹脂の流れを数値的に解析し、シミュレーションするシステムを開発したが、その解析ソフトウエア「ＦＬＯＷＡＮＡＬＹＳＩＳｏｆＲＵＮＮＥＲ」の技術移転講習会を同センターで開いた。

この直観と、同一指示の問題が関連しているかどうか確認する。

4.1 同時並行性の分析―同一指示の観点から―

3.1節で述べた複文データから、接続表現をはさんで形態素が5個以上離れたところにある動詞の組み合わせを準備調査として取り出したところ、「ため」（4210件）「が」（877件）「とともに」（546件）「ほか」

(136件) が高頻度で表れた。そこで、因果性の強い「ため」を除外し、事象を併記するという点で共通する接続表現を対象とする。

接続表現の同時並行性について、「同時並行性の違いにより事象間における同一指示が異なる」という仮説を立てる。これに基づき、下記の項目に照合して「とともに」「ほか」「が」の適用傾向を見た。

1) 後件の述語が目的語（対象）を必要とするにもかかわらず表れていなければ、前件に表れる目的語と同一指示である
2) 前件のあたまに、「同」「同様」「この」「その」がある場合、これらの表現に後接する語は、前件の主体や対象としてすでに表れている。
3) 前件のあたまに「これ」「それ」がある場合、前件全体をを指す。

上記のテストによって、「とともに」では約３０％、「が」では約１０％、「ほか」では約２％の同一指示を観察できた。とくに「とともに」では他に比べて１と２の適用度が高い。当初の予想と異なり、「ほか」より「が」の適用度が大きかったのは、下記に示すとおり「が」が逆接としてはたらく経済収支の記事中「同年比」「同期」などの同一指示が多かったためである。

例）興亜石油（社長瀬川雅夫氏）の九六年九月中間決算は、売上高は販売数量が減少したが、、販売価格の上昇で前年同期比一一・二％増の九百五十六億七千九百万円となった。

この結果は、接続表現の同時並行性の違いを示している。同一指示があること、すなわち事象間に共有する情報があることは、事象の同時性や連続性が強いと考えられる。同一指示表現によって示された事象の属性値は、接続表現の違いに着目することで特定できる可能性がある。

5．おわりに

情報抽出における事象判定の観点から、複文の分析を試みた。複文の前件・後件に表れる述語を中心に判断する各事象と、接続表現の共起パタンによって判定規則作成が可能であるとわかった。また、事象の併記という似た機能を持つ接続表現が、同時並行性の程度差によって同一指示の表れ方に相違のあることがわかった。これは情報抽出の際の同一指示判定に有効である。

本稿でも触れたが、今後の課題としては分析結果に網羅性をもたせるため現在扱えていないデータを増やすことである。また複文には、接続表現によって事象間の関係が明示的にされていない中止形の文がある。現在進めている分析から条件を見つけ出し、これらの関係を推定することも今後の課題である。

謝辞　本研究をまとめるにあたり、記事データを利用させていただいた日刊工業新聞社に感謝いたします。

また、本稿をまとめるにあたって貴重なご意見をくださった神戸市外国語大学大学院の丸山岳彦氏、ならびに九州工業大学情報工学部人間科学講座の高木一広氏に感謝申し上げます。

参考文献

[1]寺村秀夫：『日本語のシンタクスと意意味Ⅱ』くろしお出版,1984

[2]益岡隆志,田窪行則：『日本語基礎文法』「副詞節」くろしお出版,1989

[3]中川裕志：複文の意味論　因果関係を表す接続助詞を手掛りに，月刊「言語」Vol.24, No.11, pp. 46 - 53. Nov,1995

[4]益岡隆志：『複文』くろしお出版,1997

[5]西野文人，落谷亮，木田敦子，乾裕子，桑畑和佳子、橋本三奈子：トップダウンなパターン解析に基づく情報抽出，1998

[6]桑畑和佳子，橋本三奈子，木田敦子，落谷亮，西野文人：新聞記事を対象とした企業動向に関する事象構造の抽出，言語処理学会題４回念年次大会発表論文集，pp634-637,1998

[7]木田敦子，乾裕子，桑畑和佳子，橋本三奈子，落谷亮，西野文人：情報抽出のための新聞記事テキスト分析，言語処理学会題４回念年次大会発表論文集，pp238-241，1998